DOCUMENT RESUME

ED 401 329 TM 025 954

AUTHOR Marsh, Herbert A.; And Others
TITLE Is More Ever Too Much: The Number of Indicators per Factor in Confirmatory Factor Analysis.
PUB DATE 23 Jun 95
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Factor Structure; Monte Carlo Methods; *Sample Size; Simulation
IDENTIFIERS *Confirmatory Factor Analysis; *Indicators; Item Parameters

ABSTRACT

Whether "more is ever too much" for the number of indicators (p) per factor (p/f) in confirmatory factor analysis (CFA) was studied by varying sample size (N) from 50 to 1,000 and p/f from 2 to 12 items per factor in 30,000 Monte Carlo simulations. For all sample sizes, solution behavior steadily improved (more proper solutions and more accurate parameter estimates) with increasing p/f. There was a compensatory relation between N and p/f; large p/f compensated for small N and large N compensated for small p/f, but large N and large p/f was best. A bias in the behavior of the chi square was also demonstrated where apparent fit declined with increasing p/f rations even though the models were all "true." Fit was similar for proper and improper solutions, as were parameter estimates from improper solutions not involving offending estimates. The 12-p/f data were also used to construct 2, 3, 4, or 6 parcels of items (e.g., 2 parcels of 6 items per factor, 3 parcels of 4 items per factor, etc.), but the 12-indicator (nonparceled) solutions were somewhat better behaved. The study shows that traditional "rules" implying fewer indicators should be used for smaller N may be inappropriate and that CFA researchers should use more indicators per factor than is evident in current practice. (Contains 4 figures, 5 tables, and 41 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 329

Is More Ever Too Much: The Number Of Indicators Per Factor In Confirmatory Factor Analysis

Herbert W. Marsh, University of Western Sydney, Macarthur

Kit-Tai Hau, The Chinese University of Hong Kong

John R. Balla, University of Sydney

14 December, 1994

Revised: 23 June, 1995

Running Head: Is More Ever Too Much

Acknowledgments: We would acknowledge helpful comments to earlier versions of this article by Lawrence Roche, David Grayson, John Hattie, and Wayne Velicer and to thank Dennis Hocevar for stimulating some of the ideas pursued in this article. Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Faculty of Education, University of Western Sydney at Macarthur, PO Box 555, Campbelltown, New South Wales, Australia, 2560.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

HERBERT MARSH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM025954

BEST COPY AVAILABLE

ABSTRACT

We evaluated whether "more is ever too much" for the number of indicators (p) per factor (p/f) in CFA by varying sample size (N, 50-1000) and p/f (2-12 items per factor) in 30,000 Monte Carlo solutions. For all N, solution behavior steadily improved (more proper solutions, more accurate parameter estimates) with increasing p/f. There was a compensatory relation between N and p/f: large p/f compensated for small N and large N compensated for small p/f, but large-N and large-p/f was best. A bias in the behavior of the χ^2 was also demonstrated where apparent fit declined with increasing p/f ratios even though all models were "true." Fit was similar for proper and improper solutions, as were parameter estimates from improper solutions not involving offending estimates. We also used the 12-p/f data to construct 2, 3, 4, or 6 parcels of items(e.g., two parcels of 6 items per factor, three parcels of 4 items per factor, etc.), but the 12-indicator (nonparceled) solutions were somewhat better behaved. Our study shows that traditional "rules" implying fewer indicators should be used for smaller N may be inappropriate and that CFA researchers should use more indicators per factor than is evident in current practice.

Is More Ever too Much: The number of indicators per factor in Confirmatory Factor Analysis

In confirmatory factor analysis (CFA) and structural equation modelling (SEM) there is considerable confusion about the optimal amount of data that is needed to fit a given model. Applied researchers are given vague and sometimes contradictory guidelines or rules of thumb about how much data is desirable. The number of data points is the product of the sample size (N) and the number of indicators (p). Whereas there is considerable disagreement about what minimum N is desirable in CFA studies, there seems to be general agreement that more is better. In the case of the optimal ratio of number of indicators per factor (p/f), however, there is considerable disagreement about both the recommended maximum and minimum. A minimum of 3 indicators per factor is typically recommended, based in part on the classic Anderson and Rubin (1956) demonstration that this is the lower limit for identifiability. However, so long as there are multiple factors and the factors are not independent, solutions with two indicators per factor are identified and many studies are based on only two (see discussion by Bollen, 1989). Many rules of thumb imply that researchers should limit the number of indicators to be considered when N is small (e.g., rules about the minimum ratio of N/p or N/number of parameter estimates). In factor analysis there is a long history of recommendations about the minimum N/p ratio that is needed. For example, Nunnally (1967, p. 355) offered the widely cited recommendation that "a good rule is to have at least ten times as many subjects as variables." However, Tanaka (1987) argued that the ratio of N to the number of estimated parameters (t) should be more important than ratios based on the number of measured variables. Consistent with this recommendation, Bollen (1989, p. 268) stated that "though I know of no hard and fast rule, a useful suggestion is to have at least several cases per free parameter" and Bentler (1989, p. 6) suggested an "over-simplified guideline" might be that a 5:1 ratio of sample size to number of free parameters is needed when data are appropriately distributed. Because the number of variables and the number of estimated parameters tend to be substantially related, all these guidelines imply that researchers should avoid considering large numbers of indicators or estimated parameters unless N is extremely large. By implication, this means that researchers should limit p/f, particularly when N is small. In direct opposition to these implications, classical test theory suggests that it is always better to have more indicators per factor in that reliability and thus validity tend to increase.

When researchers have a large number of items per factor, a common strategy is to group the items designed to measure the same factor into 3 or more "parcels" such that each parcel is the mean of several items and then to conduct analyses on the parcel scores. Thus, for example, if there are 12 items designed to reflect a particular factor, the researcher may construct three parcels, each consisting of four items. This could be accomplished by dividing the items into three "parcels" and then computing the mean response to the four items within each parcel. The use of parcels, although not the major focus of this study, is a possible compromise between having a large number of items per factor and having a small number of indicators (parcels) in the actual analysis. This strategy is common in factor analyses of responses to rating scale items (e.g., personality tests, attitude surveys) and ability/achievement test items where it is common to analyze total test scores (averaged across a possibly large number of items) from many different tests. Although the use of

parcels is wide spread, there is surprising little systematic evaluation of its efficacy, particularly in the context of CFA. A detailed evaluation of this strategy is beyond the scope of the present investigation (but see Cattell, 1978; Comrey, 1970, 1988; Marsh, 1988; Marsh & O'Niell, 1984), but one advantage claimed for this strategy is that it decreases the number of indicators and estimated parameters relative to N. This advantage, however, is predicated on the assumption that it is good to have fewer indicators at least when N is small. The results of the present investigation may inform the appropriateness of this assumption. Also, given a sufficiently large number of items, it is possible to have very few or very many parcels. Hence, our question "is more ever too much" can also be directed at the number of parcels. More specifically, given a fixed and sufficiently large number of measured items, is it better to have 2, 3, 4, or more parcels per factor or is it better to analyze items instead of parcels?

In a possibly over-simplified form, the typical N/p and N/t guidelines for CFA imply that "more may be too much" in terms of the number of indicators per factor, whereas classical test theory implies that "more is never too much." The practical issue addressed here is to establish which of these opposing dictums is more appropriate and to explore the limits of such generalizations. More specifically, the purpose of our Monte Carlo simulation study is to determine how systematic variation in p/f and N influence the behavior of CFA solutions based on a wide range of criteria including likelihood of nonconverged solutions, the occurrence of improper solutions, interpretability, accuracy of parameter estimates, sampling fluctuations, and goodness of fit. We begin by briefly summarizing some relevant research in this area.

Tests of Statistical Significance, Convergence, and Proper Solutions

CFA is based on a sample covariance matrix S with elements s_{ij} based on sample size N and p measured variables (for more detailed descriptions see Bentler & Bonett, 1980; Bollen, 1989; Joreskog & Sorbom, 1988). It is hypothesized that the corresponding population correlation matrix Σ with elements σ_{ij} is generated by q true but unknown parameter estimates that can be expressed as a function of a q x 1 vector Θ. Thus $\sigma_{ij} = f(\Theta)$ is a model of the covariance structure where f relates the parameters in Θ to the elements σ_{ij}. Because Σ and Θ are unknown, it is necessary to estimate population parameters, resulting in Θ_E and Σ_E such that $\sigma_E = f(\Theta_E)$. The issue of goodness of fit is to determine whether S and Σ_E are sufficiently close to justify the claim that the model used to generate Σ_E fits the data. A number of different fitting functions can be used to minimize this difference, but we consider the maximum likelihood function (F_{ML}) that is the most widely used function in CFA studies. Under appropriate conditions, $(N-1)F_{ML}$ is approximately distributed as the χ^2 test statistic with $p(p+1)/2-t$ degrees of freedom, where t is the number of parameters in the model and can be used for evaluating the statistical significance of the lack of fit. Bollen (1989, pp. 266-269) noted four important assumptions for the legitimate use of χ^2 estimate: (1) the variables are multivariate normal; (2) analysis is based on the covariance matrix rather than the correlation matrix (see also Cudeck, 1989); (3) N has to be sufficiently large; and (4) the model being tested is true.

In CFA, iterative processes are used to minimize the difference between S and Σ_E in relation to a particular fitting function. This iterative process continues until the difference between any two steps is smaller

than some pre-determined value, a criterion of convergence. Nonconvergence occurs when the estimation algorithm is unable to meet the criterion within a specified number of iterations (see Bollen, 1989, p.254; Joreskog & Sorbom, 1988, p.269). In addition to the characteristics of the model and the data, convergence could be dependent on the number of iterations allowed, the criterion of convergence, and, possibly, the starting values used in the first step of the iterative process. In LISREL, for example, the default criterion of convergence is set at a value that generally provides parameter estimates accurate to three significant digits whereas the number of iterations is set to three times the number of parameters in the models. Joreskog and Sorbom suggested that, "Our experience is that, for models which are reasonable for the data, the iterations will converge before this maximum is reached" (1988, p.182) so that nonconvergence cannot be solved simply by increasing the number of iterations allowed. Whereas there is some possibility that nonconvergence may depend on the initial starting values for the iterative process, Boomsma (1985) found nearly identical solutions using a variety of different starting values and final solutions are rarely reported to vary depending on the starting values (see Marsh, Byrne & Craven, 1992). Hence, it seems that the problem of nonconvergence is more likely to be a function of the data or the model rather than the number of iterations or starting values, a conclusion that is consistent with the position advocated by Velicer and Jackson (1990).

Even when there is convergence, the obtained solution may not be interpretable. Of particular relevance to the present investigation, the solution may be improper such that one of more of the parameter estimation matrices is not positive definite (Bentler & Jamshidian, 1994; Wothke, 1993). Thus, for example, a variance or residual variance estimate may be negative (typically called a Heywood case) or a factor correlation (standardized factor covariance) may have an absolute value greater than 1.0. van Driel (1978; see also Bollen, 1989; Dillon, Kumar, & Mulani, 1987) recommended that the formal requirement of positive definiteness be dropped, thus allowing researchers to distinguish between three classifications of improper solution: (1) boundary cases in which the confidence interval around the offending parameter contains proper values (e.g., the confidence interval around a negative uniqueness includes positive values) so that the problem may merely reflect sampling fluctuations; (2) nonboundary cases in which the confidence interval for the offending parameter does not contain any proper values; and (3) indefinite cases where the standard error (SE) is so large that no interpretations are warranted even though the confidence interval may contain proper values (called large SE solutions in the present investigation). van Driel suggested the existence of boundary cases may not require rejection of the hypothesis of an interpretable factor structure, but that nonboundary and large SE solutions may require researchers to delete existing variables, add new variables, or respecify the model.

Boomsma (1985) and Gerbing and Anderson (1987) compared χ^2 estimates and goodness of fit indices for fully proper and converged-improper solutions, but found little systematic differences. However, Boomsma (1985) did find that parameter estimates from solutions containing improper solutions tended to be somewhat more biased and to have somewhat larger standard errors than those that excluded improper cases. Considering all aspects of his research, Boomsma concluded that it was still disputable as to whether or not to include improper solutions in Monte Carlo studies but reiterated that researchers should use Ns of at least 100. Gerbing

and Anderson added further clarification about the interpretability of improper solutions by categorizing parameter estimates from improper solutions into three categories: (a) offending parameters (e.g., Heywood cases and associated factor loadings); (b) offending-related parameters (parameter estimates for measured variables in the same factor as offending parameters); and (c) non-offending parameters (estimates not involved with factors having offending parameters). Like Boomsma (1985), they found that parameter estimates in improper solutions varied from those in proper solutions. Offending uniquenesses were negative and the corresponding factor loadings were positively biased, whereas offending-related parameter estimates were biased in the opposite, compensatory direction (e.g., factor loadings were negatively biased). Their important new finding, however, was that non-offending parameter estimates in improper solutions did not differ significantly from parameter estimates in fully proper solutions. Hence, even when the solution is improper, it appears that many of the parameter estimates are interpretable.

Effects Of Sample Size (N) And Number Of Indicators Per Factor (P/F)

Sample Size.

Much of the relevant literature in this area is summarized by the Gerbing and Anderson (1993) chapter on Monte Carlo evaluations of goodness of fit. In his classic Monte Carlo study, Boomsma (1982) evaluated the robustness of CFA solutions for small Ns (25 to 400). He found that the percentage of proper solutions, accuracy of parameter estimates, sampling variability in parameter estimates, and the appropriateness of the χ^2 test statistic were all favorably influenced by having larger Ns. Based on this research, Boomsma offered his widely cited recommendation that N should be at least 100, but also noted that Ns of 200 or more may be desirable in some circumstances. For each level of N, solutions were better behaved for p/f = 4 than for p/f = 2 when the saturation of measured variables (the relation between a measured variable and its latent factor, the factor loading) was larger. The main finding that larger Ns were associated with better behaved solutions has been replicated in many subsequent studies (e.g., Anderson & Gerbing, 1984; Boomsma, 1985; Gerbing & Anderson, 1987; 1993; also see Guadagnoli & Velicer, 1988).

Velicer and Fava (1987, 1994) argued that concerns about the minimum N for factor analysis have produced many guidelines but limited empirical research. They briefly reviewed recommendations based on an absolute minimum N (e.g., 100 or 200) and ratios of N/p ranging from 2 to 20, but noted that the most familiar advice was to have as large an N as possible. They concluded that there was no support for rules positing a minimum N as a function of p. In fact, the Guadagnoli and Velicer (1988) principal component study found that for a fixed N better results were obtained if p/f was larger, not smaller. Velicer and Fava found that convergence to proper solutions and goodness of fit were favorably influenced by increasing N, p/f, and saturation (factor loadings), and that these results were similar for principal component analysis, image component analysis, and maximum likelihood factor analysis. One purpose of the present investigation is to replicate and extent these results in terms of N and p/f for CFA.

Marsh and Bailey (1991) evaluated the behavior of a set of models designed to fit "real" and simulated multitrait-multimethod (MTMM) data. In a typical MTMM model, T x M measured variables are used to infer

T correlated trait factors and M correlated method factors. Unlike most Monte Carlo studies in which each measured variable reflects only one latent construct, in MTMM studies each measured variable is posited to reflect one trait factor and one method factor. Across 435 MTMM matrices, this model converged to a proper solution only 24% of the time. The likelihood of convergence to proper solutions increased as N and p increased. However, even when MTMM data was simulated from this MTMM model, most of the solutions were improper when N and p were small (e.g., N = 160 or 400 and p = 9, a 3Tx3M design). Proper solutions were much more likely when N and particularly p were large (e.g., p= 36 in a 6T x 6M design, or p = 28 in a 7T x 4M design). The authors noted that their results contradicted rules of thumb about N/p, and also called into question the generalizability of claims that three indicators per factor are sufficient to produce stable, well-defined structures. (Note that the number of indicators per factor is somewhat ambiguous in MTMM studies where each measured variable loads on two factors. For example, in a 3T x 3M design, each latent factor is inferred from three indicators but a total of 6 factors are inferred from only 9 measured variables.) Whereas MTMM data is not the focus of the present investigation, this research demonstrates that the generalizability of simple rules of thumb may be limited.

In summary, the only recommendation about N that has received consistent support is the claim that more is better. In particular, there appear to be no generalizable guidelines about the minimum N needed to achieve a stable, well-defined solution. Rules positing minimum ratios of N to p or t seem to be violated by Velicer and Fava. (1987, 1994; also see Guadagnoli & Velicer, 1988) factor analysis studies and the Marsh and Bailey (1991) CFA study of MTMM data. In the present investigation, we test the generalizability of these results in a CFA study in which an "independent clusters" model (i.e., each measured variable loads on one and only one factor) serves as both the generating model used to create the simulated data and the approximating model used to fit the simulated data.

Number Of Items Per Factor

A few studies have systematically evaluated the effects of p/f in conjunction with other variables, and reasonably systematic effects have been found. Thus, for example, Anderson and Gerbing (1984), Boomsma (1982), Boomsma (1985), Ding, Velicer, and Harlow, (in press; also see Ding, 1993) and Gerbing and Anderson (1987) reported that the likelihood of fully proper solutions increased with increasing p/f, N, and saturation. Similar results were reported in Monte Carlo studies involving principal components and factor analysis (Velicer & Fava, 1987). Gerbing and Anderson (1987) also demonstrated that the standard errors of parameter estimates were smaller when N and p/f were larger.

Anderson and Gerbing (1984), Boomsma (1982), Bearden, Sharma and Teel (1982), Ding et al. (1994; also see Ding , 1993), and Gerbing and Anderson (1987) also reported, however, that goodness of fit tended to be negatively related to increasing p. In an early Monte Carlo study of 2- and 4-factor models (both having 3 items per factor), Bearden et al. (1982) reported that the χ^2 test behaved appropriately for two-factor models, but that for four-factor models with small N the test statistic led to too many rejections of the null hypothesis. These results suggest that p may be critical, although p/f ratios were not manipulated independently of p.

Anderson and Gerbing (1984, p.167) reported mean p-values for the χ^2 test of .536, .461, and .416 (with SEs for the mean of less than .01 due to the large number of replicates in their study) for p/f = 2, 3, and 4 respectively. All pair-wise differences were significant, the value for p/f = 2 was significantly higher than the expected value of .5 (i.e., a bias in the direction of too few rejections), and the values for p/f = 3 and 4 were significantly smaller than expected. This effect was larger for small N and was attenuated for large N. Boomsma (1982, p. 168) also reported χ^2 s to be slightly larger than expected for p/f=3 and particularly p/f=4 when N was small and factor loadings were moderate or large. A similar pattern was also reported by Ding et al (in press; also see Ding, 1993). They found that the probability of rejecting true models (at $p < .05$) was close to 5% for p/f =2 but rose steadily as p/f ratios increased. This effect varied with N so that for p/f=6, rejection rates were 39% for N=50, 22% for N=100, 12% for N=200, and 6% for N=400. The size of this effect was more dramatic in the Ding et al study than previous research because they considered a wider range of p/f ratios (2 to 6 indicators per factor) than the other studies summarized here. Because all approximating models in each of these studies were true models, these results suggest a systematic bias in the χ^2 that varies as a function of p/f.

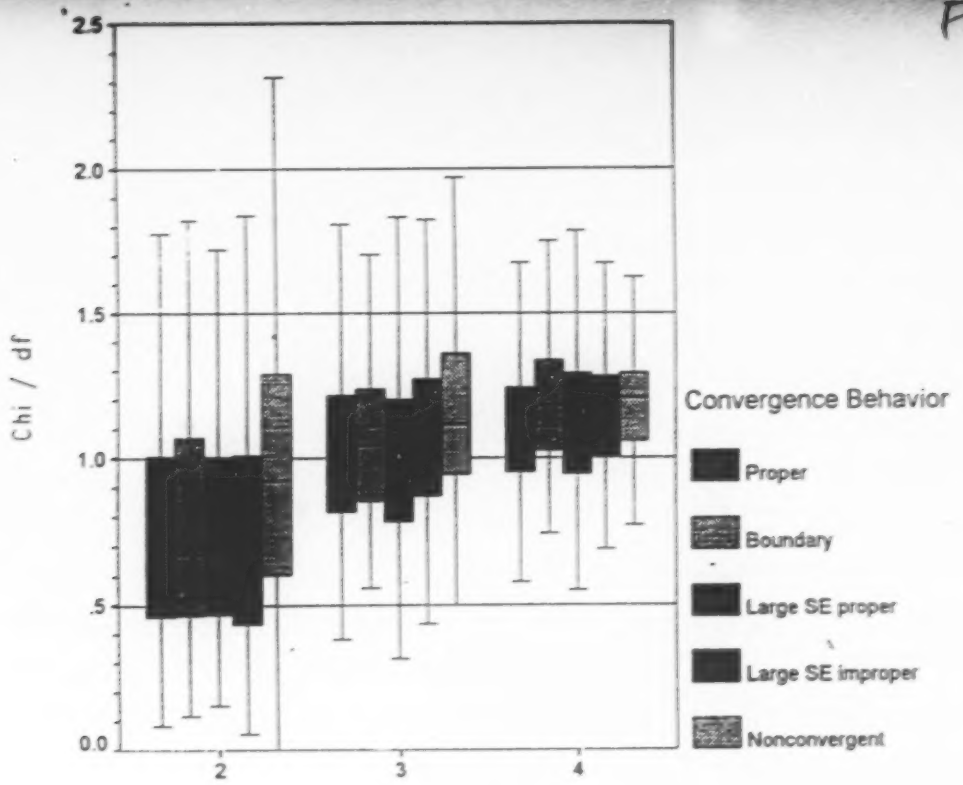
Methodology Overview

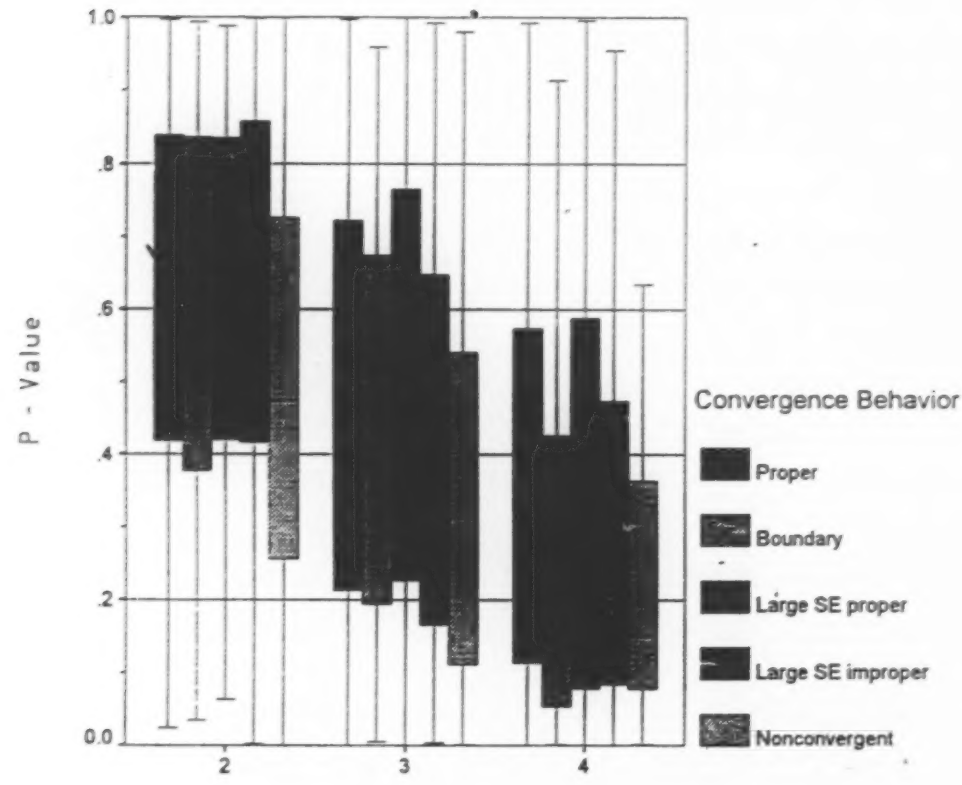
The present investigation consists of three Monte Carlo simulation studies using a largely parallel methodology and overlapping data. All analyses for the present investigation were conducted with the PC version of LISREL 8 (Joreskog & Sorbom, 1993) and version 6 of SPSS for Windows (SPSS, 1993). Three-factor congeneric models (see Figure 1) were constructed in which each indicator loaded on one and only one factor (see Marsh, Balla & McDonald, 1988; Marsh & Balla, 1994). Five levels of p/f were considered in which all factors for a given generating model had 2, 3 (as shown in Figure 1), 4, 6, or 12 items per factor. All factors had unit variance and were correlated .30 with each other. Each indicator also had unit variance and had factor loadings and uniqueness of .60 and .64 respectively. Five levels of N (number of cases per replicate) were 50, 100, 200, 400, and 1000. For purposes of this study, the simulated data were generated using the random generator function NORMAL in SPSS. In order to evaluate the effects of N, these cases were divided into 2500 replicates of N=50 cases, 1000 replicates of N=100, 500 replicates of N=200, 250 replicates of N=400, and 100 replicates of N=1000. We specifically generated more replicates for smaller N because the behavior of the small N solutions is a primary focus of this investigation and these solutions tend to be less stable.

Insert Figure 1 about here

For all LISREL analyses in the present investigation: the factor variances were fixed at unity; factor correlations, factor loadings, and uniqueness were all freely estimated; the maximum number of iterations was set to 500; LISREL's default starting values were used; and the default check on the admissibility of the solution was turned off. All models tested in this study are "true" in that the pattern of free and fixed parameters was the same in the generating model used to generate the simulated data and the approximating model used to fit the data. Hence, to the extent that the χ^2 test is behaving appropriately, the mean χ^2 /df ratio should not differ systematically from 1.0 and the mean p-value associated with the χ^2 should not differ significantly from .50. (Because the df associated with different models varies substantially, the χ^2 should differ substantially from

Fig 1





Number of Indicators per Factor

model to model and so it is less useful for present purposes). Because the purpose of the present investigation was not to evaluate different fit indices per se, we do not present results for a variety of subjective indices of fit (e.g., Marsh et al., 1988; McDonald & Marsh, 1990), but we briefly summarize these findings in relation to the χ^2 /df results.

Study 1: Nonconverged and Improper Solutions

Study 1 focuses on the likelihood that solutions are fully proper, improper, or nonconverged. Consistent with previous research, preliminary analyses demonstrated that nonconverged and improper solutions are more likely when N and p/f were small. Because of the focus of study 1, we evaluated data that produced many improper solutions (N=50 and p/f = 2, 3, or 4) by comparing goodness of fit and parameter estimates from fully proper solutions with those from various classifications of improper solutions.

For purposes of comparison, the solutions were divided into six categories following the classification by van Driel (1978): proper, boundary, non-boundary, large SE proper, large SE improper, and nonconverged. The boundary and nonboundary cases consisted of solutions with offending parameters (i.e., negative uniqueness or factor correlations greater than 1.0 in absolute value). The boundary and non-boundary cases differed in that offending parameters of the former were within 2 standard errors (estimated from fully proper solutions) of the permissible region boundaries (0 for uniquenesses and ± 1 for correlations). Large SE solutions were operationally defined as all solutions having at least one estimated SE that was larger than the corresponding mean SE of the fully proper solution by at least 5 SEs. These large SE solutions were further divided into large SE-proper and large SE-improper, depending on whether there were any offending parameter estimates in addition to the large SE. Thus, large SE-proper solutions had at least one excessively large SE but no negative uniquenesses and no factor correlations greater than 1 in absolute value. Nonconverged solutions were operationally defined as the failure to converge to the default LISREL criterion of convergence within 500 iterations.

Following Gerbing and Anderson (1987), the factor loadings, uniquenesses, and factor correlations were classified as proper, offending, offending-related, and non-offending. Proper parameter estimates were all those in fully proper solutions. Offending parameters referred to negative uniquenesses (and the corresponding factor loading), factor correlations greater than 1 in absolute value, and parameter estimates with excessively large SEs. Offending-related parameters were factor loadings, uniquenesses, or factor correlations associated with a factor with an offending parameter estimate. Non-offending parameter estimates were all remaining parameter estimates in an improper solution. Thus, for example, if the uniqueness for the first indicator of factor 1 was negative (or had an excessively large SE), then factor loadings and uniquenesses for that indicator were classified as "offending" estimates; all factor loadings and uniquenesses for other indicators of factor 1 and all factor correlations involving factor 1 were classified as "offending-related" estimates; and all factor loadings and uniqueness for indicators not loading on factor 1 and factor correlations not involving factor 1 were classified as "non-offending" estimates.

In summary, Study 1 examines the effects of p/f on the behavior of CFA solutions, following from and extending the earlier research (Gerbing & Anderson, 1987; Dillon, et al, 1987; van Driel, 1978; Velicer & Fava, 1987, 1994). In particular, we examine various classifications of improper solutions and make separate comparisons of offending, offending-related, and non-offending parameter estimates. More importantly, we bring together these two themes by comparing different types of parameter estimates in the various classifications of improper solutions.

Results: Nonconverged and Improper Solutions (Study 1)

Convergence To Proper Solutions. The effects of N and p/f ratios on the convergence are examined first (Table 1, ignoring the results for parcels for now). The likelihood of fully proper solutions is substantially related to larger Ns and larger p/f. Thus, for example, only 13.6% of the solutions based on N=50 and p/f = 2 are proper (56.6% of the solutions were nonconverged), whereas solutions based on N=50 and p/f=12 or N=1000 and p/f=2 both converged to proper solutions most of the time (100% & 93% respectively). These results suggest a compensatory effect of N and p/f on the behavior of solutions. For p/f = 2, it is important to have very large Ns -- at least N=400 and preferably more for data considered here -- in order to be reasonably confident in obtaining a fully proper solution. When p/f = 3 (the minimum typically recommended), the results are consistent with Boomsma's widely cited recommendations in that N=100 may be sufficient but that N=200 is preferable. However, when p/f = 6 or 12, N=50 is sufficient. These results support a "more is better" conclusion for both N and p/f. Whereas Boomsma's (1982) recommendation of a minimum N=100 seems reasonable when p/f = 3 or 4, it does not generalize to solutions where p/f is smaller (p/f = 2) or larger (p/f= 6, or 12). Likewise, whereas it may be undesirable to have p/f = 2, this situation is less unacceptable when N is sufficiently large. Finally, the results offer a strong refutation of guidelines focusing on minimum ratios of N to p or t. At any given N in Table 1 the likelihood of convergence to a proper solution improves with increasing p/f.

Insert Table 1 about here

Parameter Estimates. Parameter estimates from fully proper, nonconverged, and improper solutions (Tables 2 and 3) are based on the 2,500 replicates for cells with N=50 and p/f = 2, 3, or 4 where the occurrence of improper and nonconverged solutions is frequent. All models considered in the present investigation have at least 2 indicators for each of the three factors and so a total of six factor loadings and six uniquenesses are considered for each solution (Table 2). Hence the total number of factor loadings (and uniquenesses) for each of the three cells considered here is 15,000 (6 estimates x 2500 replicates). Similarly, because all models have three correlations, the number of correlations summarized in each cell is 7,500 (3 estimates x 2,500 replicates). The classification of the solution type -- proper, nonconverged, etc. -- was, of course, based on the entire set of parameter estimates from each solution.

Insert Table 2 about here

The critical comparisons in Table 2 are those between parameter estimates from fully proper solutions and those from the various categories of improper solutions. For all three cells of the design, non-offending

parameter estimates from improper solutions are reasonably similar to the corresponding parameter estimates from fully proper solutions. For p/f = 2, the nonconverged/nonoffending factor loadings are substantially lower and more variable than the fully proper estimates, but otherwise even the nonoffending parameter estimates in the nonconverged cases are reasonable. Particularly when there are at least three items per factor, all nonoffending parameter estimates from nonconverged and different types of improper solutions were similar to those based on the fully proper solutions.

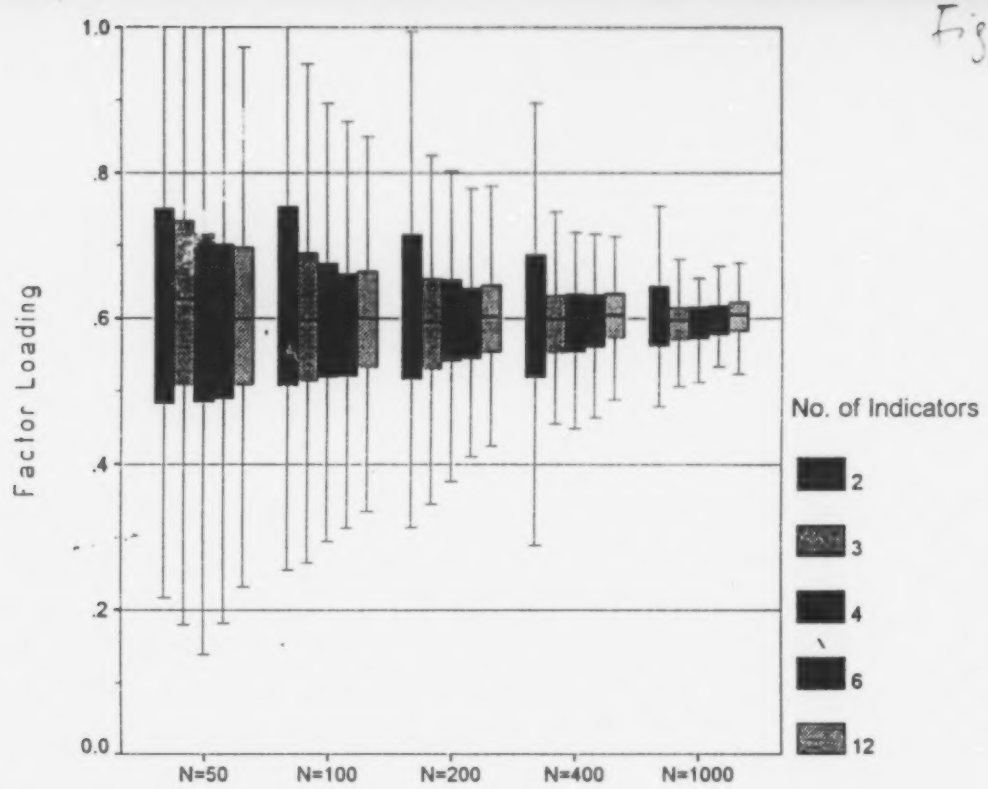
There is also a clear and consistent pattern for the offending and offending-related estimates of factor loadings and uniquenesses. For all cells the offending uniquenesses are consistently much too small, whereas the corresponding factor loadings are much too large. Whereas these results for the offending parameter estimates are not surprising, it is important to emphasize that for large SE solutions that were otherwise proper, the offending parameter estimates (i.e, otherwise proper estimates with large SEs) are also systematically biased. Somewhat more surprising, though still consistent with results reported by Gerbing and Anderson (1987), is the pattern of offending-related factor loadings and uniquenesses. In all cells the offending-related uniquenesses are substantially larger than in the fully proper solutions, whereas the offending-related factor loadings are substantially smaller. There is not such a clear pattern of results for factor correlations. Whereas non-offending factor correlations appear to be reasonably similar to those based on fully proper solutions, other factor correlations from improper solutions are not.

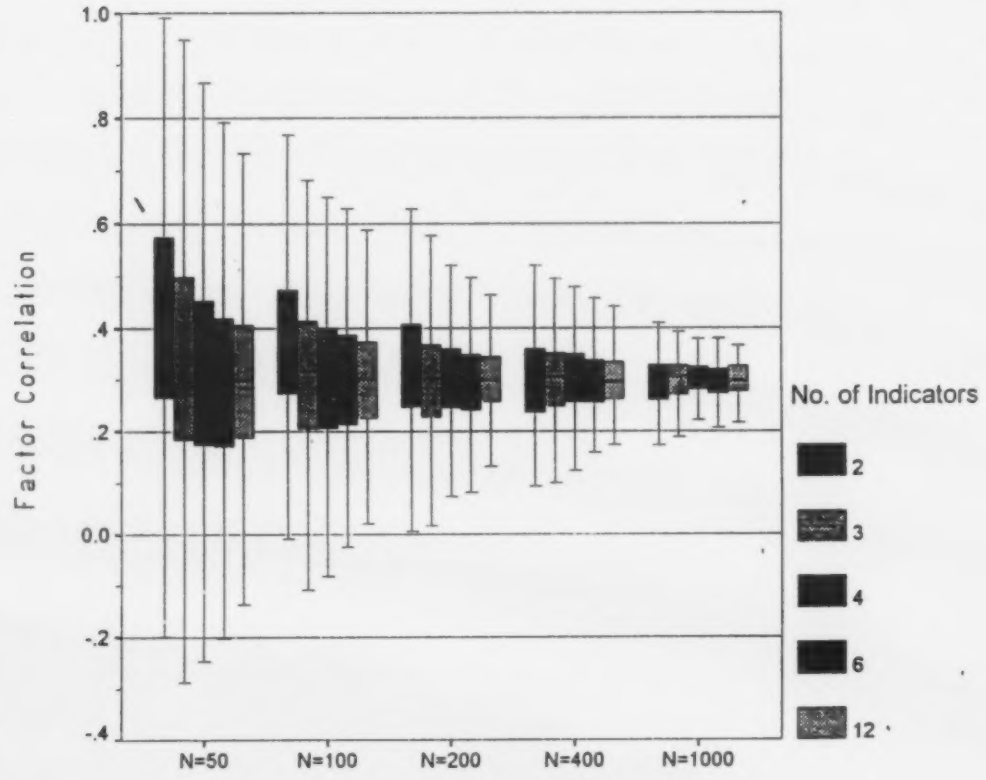
Goodness of Fit. The means and SDs for the goodness of fit indices are shown in Table 3. Group means were compared with oneway ANOVA followed by post-hoc Scheffe tests (SPSS, 1993). Despite the large number of cases (2500 replicates), the goodness of fit statistics for the fully proper solutions do not differ significantly from any classification of (converged) improper solutions in analyses summarized in Table 3(also see boxplots in Figure 2). In fact, except for the solutions based on two indicators per factor, even the goodness of fit statistics for the nonconverged solutions are reasonably similar to those from the fully proper solutions. Also, the standard deviations for the fully proper solutions are reasonably similar to those for the different categories of improper solutions and, except for the two-indicator cell, even the nonconverged solutions.

Insert Figure 2 and Table 3 about here

Although not the major emphasis of this study, it is disconcerting to note that the mean values for all the fit indices differ systematically with p/f. Consider, for example, the χ^2 /df ratio (Figure 2) that is expected to be 1.0 for all cells (since only true models are considered here). For fully proper and all categories of improper solutions, the χ^2 /df ratio is systematically smaller than 1.0 for two-indicator solutions, slightly larger than 1.0 for three-indicator solutions, and substantially larger than 1.0 for the four-indicator solutions. A naive interpretation of these results might suggest that the goodness of fit is somehow better when there are fewer indicators. However, because all the solutions are based on true models, it appears that these results reflect a breakdown in the asymptotic behavior of the χ^2 statistic. Because this issue is even more evident in Study 2 where there is a wider range of p/f and N, we revisit the implications of this finding in our overview discussion of the two studies.

Fig 2





Sample Size

Study 2: Effects of Number of Indicators and N in Confirmatory Factor Analysis

In study 2 we considered only the fully proper solutions from the 25 cells of our design representing the five levels of p/f (2, 3, 4, 6, and 12) and five levels of N (50, 100, 200, 400, and 1000). As in Study 1, fully proper solutions were operationally defined as converged solutions with no negative uniquenesses, no factor correlations greater than ± 1.0, and reasonable SEs for parameter estimates. The focus of this study was on the parameter estimates, the variability of parameter estimates, goodness of fit, and factor reliability across the 25 cells in the design. In order to summarize our results, we conducted two-way ANOVAs, but because of the very large number of replicates our primary consideration was on effect sizes instead of nominal tests of statistical significance. Effect sizes presented are eta (the square root of the ratio of $SS_{explained}/SS_{total}$ for each effect) and r (the linear effect of log N, the linear effect of log of p/f, and the linear x linear component of the N x p/f interaction).

Results: Effects of Number of Indicators and N in Confirmatory Factor Analysis (Study 2)

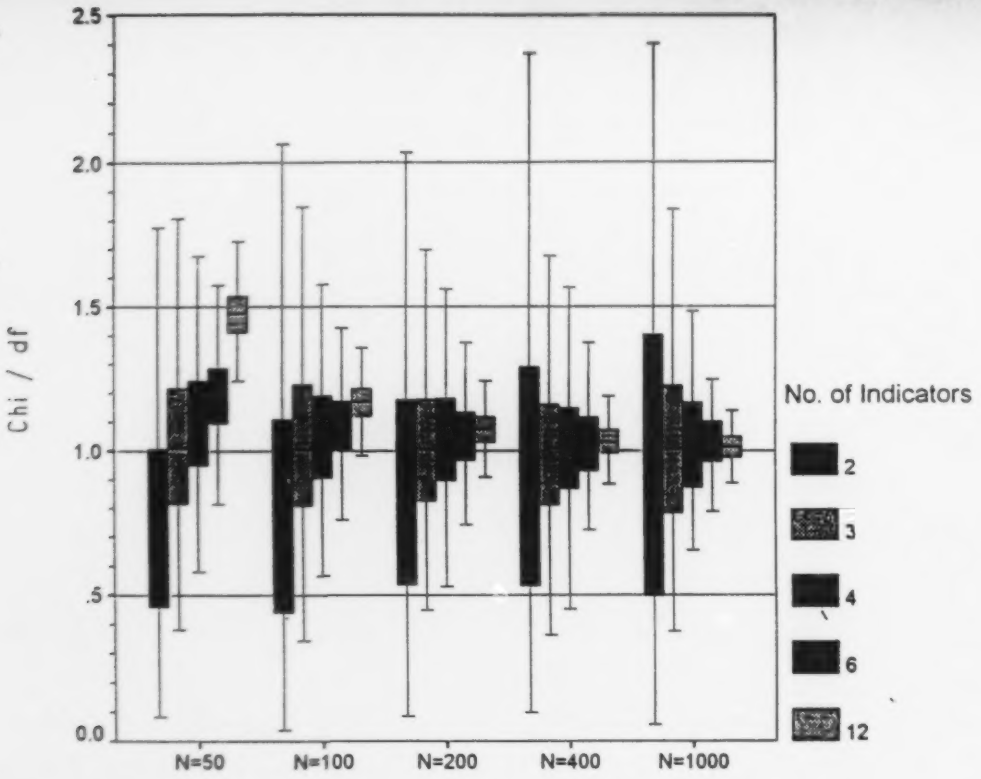
Parameter Estimates. The means and SDs of the factor loadings, uniquenesses, factor correlations, and factor reliabilities are presented in Table 4 (ignoring parcel results for now) and the effect sizes are summarized in Table 5 (also see box plots in figure 3). Mean parameter estimates for all cells are reasonably consistent although estimates from two-indicator solutions -- particularly those with smaller Ns appear to differ systematically from population values. However, a two-way ANOVA of these results (Table 5) indicated that the combined effects of p/f, N, and their interaction accounted for only 0.1%, 0.5%, and 1.4% of the variance in factor loadings, uniquenesses, and factor correlations respectively. In contrast to the mean parameter estimates, the mean factor reliability estimates increase steadily with increasing p/f.

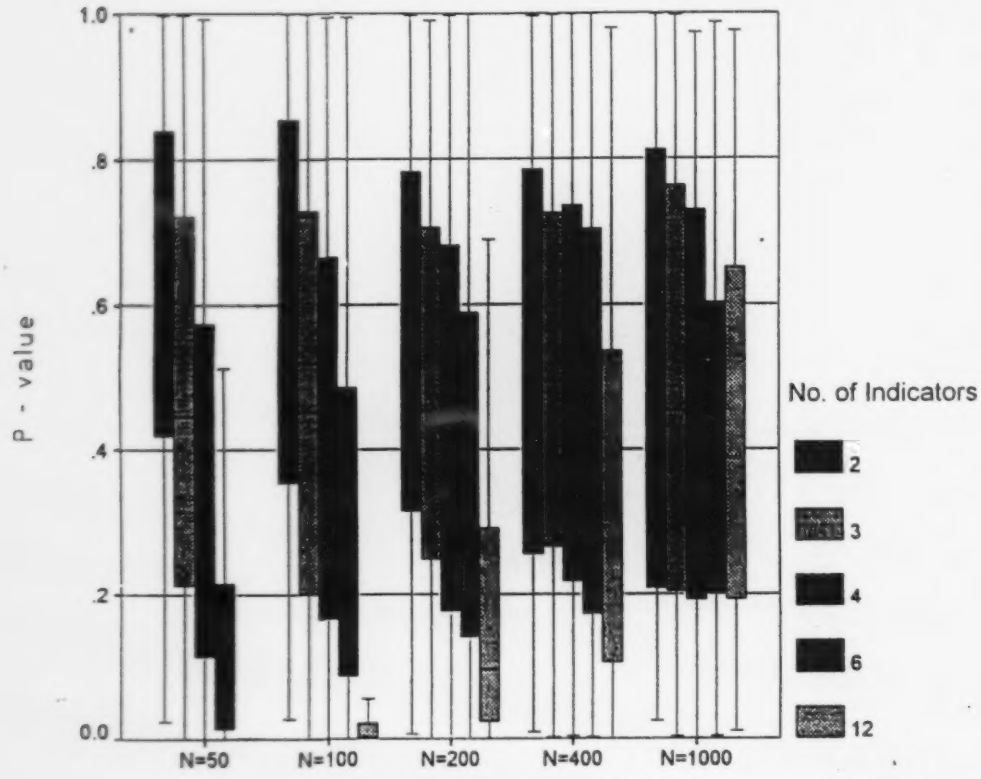
The SDs for all parameter and reliability estimates decrease substantially with increases in N and p/f (see Table 4). In order to more systematically evaluate these effects, we computed absolute difference scores for the first two factor loadings on factor 1, the first two uniquenesses for factor 1, and the first two factor correlations and related these to N, p/f, and their interaction. For these analyses, substantial portions of the variance were explained by these effects; 12.9%, 14.7%, and 14.3% of the variance in factor loadings, uniquenesses, and factor correlations respectively. Although the effects of N and p/f are large, their interaction explained less than 1% of the variance in each set of estimates (Table 5).

Insert Tables 4 and 5 and Figure 3 about here

Goodness of Fit. The means and SDs of the goodness of fit indices are presented in Table 4 and the corresponding two-way ANOVAs are summarised in Table 5 (also see boxplots in Figure 4). Because the linear effects are typically only marginally smaller than the corresponding etas, most of the explained variance can be accounted for by the (log) linear effects. Goodness of fit appears to decline as p/f increases and this effect is larger when N is small. Although a smaller and less systematic trend, goodness of fit appears to become poorer with increasing N when p/f is small and to improve with N when p/f is large. The fit indices most closely approximate their expected values (for a true model) when N =1000. As noted previously, these apparent differences in fit are illusory in that only true models are considered. Hence, these effects apparently reflect the breakdown in the asymptotic properties of the χ^2 statistic.

Fig 3





Sample Size

Study 3: A Comparison of Parcel and Item Solutions

The purpose of Study 3 is to explore the use of parcels in relation to our broader question of whether more is better. In Study 3 we extended consideration to parcels by constructing 2, 3, 4, and 6 parcels from the 12 items in the p/f = 12 data considered earlier. Thus, the 12 items per factor were divided into two parcels (of six items), three parcels (of 4 items), four parcels (of 3 items), and six parcels (of 2 items). In each case, the simple mean of the responses to the items in each parcel was used to represent the parcel score. Hence, each latent construct reflected all 12 items, but the actual number of parcels used in the CFA varied.

All analyses were conducted in the covariance metric. This is particularly important because the expected variance of each parcel score, depending on the number of items in the parcel, is substantially less than the variability of the item scores (1.0 in the population in the present investigation) used to construct the parcel by virtue of the central limit theorem. This feature of the parcels also complicates comparisons between item and parcel solutions. In addressing this issue it would be possible to compare 2-item solutions with 2-parcel solutions, 3-item solutions with 3-parcel solutions, etc. However, such comparisons are of limited value because they confound effects of the number of indicators (items or parcels), the number of items, and therelative saturation of each indicator. It is more relevant to compare the results based on the various parcel solutions with each other and with the 12-item solution, because an applied researcher with a fixed number of items per factor must decide whether to analyze items or parcels.

Results: A Comparison of Parcel and Item Solutions (Study 3)

Proper and Improper Solutions. Solutions based on 2, 3, 4 or 6 parcels are more likely to be fully proper than solutions based on 2, 3, 4, or 6 items (Table 1). Nevertheless, the trends are similar in that the likelihood of a proper solution increases systematically with N and the number of indicators (items or parcels). However, the advantage of parcels is primarily evident for the 3- and 4-indicator solutions. Although these results appear to favor parcel solutions, it is important to reiterate that all the parcel solutions were based on responses to 12 items and that the 12-item solutions resulted in 100% fully proper solutions for all N. When the number of parcels is small and N is small to moderate, parcel solutions are more likely to result in improper solutions than the 12-item solution. Hence, at least in terms of obtaining fully proper solutions for the data considered here, there are some potential disadvantages in using parcels instead of the individual items to construct the parcels. These results are consistent with earlier results suggesting that the use of more indicators is better.

Parameter Estimates and Goodness of Fit. Parameter estimates for fully proper parcel solutions are presented in Table 4 along with the item solutions considered earlier (also see effect sizes in Table 5). These comparisons, however, are substantially influenced by the change in metric. Whereas the population variance of each item is 1.0, the population variance of parcel scores decreases systematically with increases in the number of items included in each parcel (consistent with the central limit theorem). Mean factor loadings and correlations are almost unaffected by N or p/f. The large effect of the number of parcels on the uniquenesses reflects in part the metric of the parcels. For analyses done in the (untransformed) covariance metric, the factor

loadings remain constant whereas the uniquenesses (error variance) decrease as the number of items in each parcel increase. Note, however, that the reliability of each factor is nearly unaffected because all parcel solutions and the 12-item solutions are all based on responses to the same 12 items. It is also interesting to note that if the analyses had been conducted in a correlational metric (or, equivalently, reported in terms of completely standardized estimates) the size of factor loadings would systematically increase as the number of parcels decreased (and the number of items used to construct each parcel increased). However, this merely reflects the scaling of the parameter estimates and does not fundamentally affect the interpretation of results presented here.

The parameter estimate SDs all decline with increasing N as observed for the item solutions (see Tables 4 and 5). The effect of the number of parcels on the parameter estimate SDs, however, is complicated by the change in metric of the analyses. Because each parcel reflects the average of several independent measures, the variability of the factor loadings and uniquenesses vary inversely with the number of items in the parcel. This added precision associated with parcels based on larger numbers of items, however, is offset by the fact that the number of estimates is smaller. Thus, for example, the factor loadings for the 12-item solutions are much less precise than the factor loadings based on 2 parcels, but there are many more independent estimates of the factor in the 12-item solution than the 2-parcel solution. Reflecting this trade-off between the precision of estimation and the number of estimates, the SDs of the correlations are nearly independent of the number of parcels. That is, because all latent constructs are inferred from responses to all 12 original items no matter how many parcels are included in the analysis, inferences about the latent constructs are similar for item and parcel solutions. This pattern is also consistent with the finding that the mean reliability is similar across all cells for the parcel solutions. The SDs of the reliability estimates, however, do decline systematically with increasing N and p/f.

The goodness of fit of parcel solutions (Table 4) for any given Ns appear to become poorer as the numbers of parcels increase, but this pattern reflects the bias in the χ^2 observed earlier. Again, the size of this apparent bias systematically decreases with increasing N, suggesting that the rate at which the χ^2 approaches its asymptotic behavior varies with the number of parcels.

Summary and Implications

Number of Items

Our major focus has been on the question "is more ever too much" in relation to the number of indicators per factor in CFA studies. In answer to this question, the present investigation provides clear support for the advantages of using more indicators per factor: fewer nonconverged solutions, fewer improper solutions, greater interpretability (even when solutions are improper), more accurate and stable parameter estimates, and more reliable factors. This use of more indicators per factor is particularly important when N is small to moderate. Recommendations that N should be at least 100 are widely cited, but N=100 was not sufficient when p/f was only 2 or even 3 in the present investigation. With N=100, researchers should have at least four items per factor based on the model used here and more is better. Furthermore, Ns smaller than 100 may suffice if p/f is sufficiently large although more is better. In the present investigation, for example, solutions with N=50

always converged to fully proper when there were 12 items per factor. Within the limitations of this study it appears that it is better to have more indicators per factor so that more is never too much.

We do not expect that the actual values obtained in this research to generalize to other CFA models. Thus, for example, there is some evidence that the minimum desirable N and p/f may also depend on the saturation of the indicators (eg., Ding, et al., in press; Velicer & Fava, 1987, 1994). Support for this conclusion is also evident in Study 3 where the use of parcels effectively increased the relative saturation of the parcel indicators and parcel solutions were somewhat more likely to result in fully proper solutions than corresponding item solutions (e.g., 3-parcel/factor solutions vs. 3-item/factor solutions). Also, in the present investigation we only considered simple population generating models (each indicator loaded on one and only one factor) and true approximating models. As demonstrated in the Marsh and Bailey (1991) study of real and simulated MTMM data, the situation becomes more complicated when indicators load on more than one factor and real data are considered. Whereas further research is needed to clarify how many indicators per factor are needed in a much wider set of circumstances than considered here, we predict that our "more is better" conclusion will have broad generalizability.

In simulation studies it is easy to generate ever increasing numbers of high quality indicators, but in practice this may not be the case. Particularly in exploratory studies, researchers may not even know the quality of the finite number of items available. In this situation it is even more important to include a large number of items so that there will be a sufficient number of good items even if some items are subsequently discarded. In the context of factor analysis, for example, Velicer and Fava (1994) suggested that 6-10 items per factor is a good initial target, based on the assumption that 25-50% of the initial items will have to be deleted, but recommend that 20-30 items are needed when the quality of items is poor or N is small. Although these recommendations are more extreme than conditions considered in the present investigation, they are consistent with our claim that more is better.

Nonconverged and Improper Solutions

In the present investigation we extended and combined a classification of solutions (fully proper, boundary, non-boundary, large SE improper, large SE proper, nonconverged) adapted from van Driel (1978) and a classification of parameter estimates (proper, nonoffending, offending related, and offending) adapted from Gerbing and Anderson (1987). The fit of fully proper solutions did not differ substantially from values found in any of the classifications of improper solution considered here. Except for the solutions with p/f = 2, not even the fit of the nonconverged solutions was appreciably poorer than the fit of the fully proper solutions. These results are consistent with previous research (e.g., Boomsma, 1985; Gerbing & Anderson, 1987), although these studies did not specifically evaluate the fit of nonconverged solutions. These findings suggest that it may be reasonable for researchers to evaluate the goodness of fit of improper and, perhaps, even nonconverged solutions. Thus, even when one model within a nested sequence of models does not result in a fully proper solution, it may be useful to evaluate how the fit of this improper solution compares with other fully proper solutions. For example, in evaluating the partially nested taxonomy of MTMM models (Marsh, 1989),

the most general model typically does not produce fully proper solutions. However, Marsh suggested that it may still be reasonable to compare the goodness of fit of this general model with more restrictive models in the taxonomy that do converge to fully proper solutions. Particularly if the two models are nested, then this comparison may support the conclusion that the added complexity in the more complex model is unnecessary or unwarranted. This strategy may be particularly useful when the researcher posits an a priori set of fully or partially nested models instead of relying on the evaluation of a single model.

In our study, non-offending parameter estimates from improper solutions did not differ substantially from parameter estimates in fully proper solutions. This pattern of results was reasonably consistent across the different classifications of improper solution and, except for p/f=2, even the nonconverged solutions. In contrast, the offending and offending-related estimates differed systematically from those in proper solutions. Hence, it may be reasonable for researchers to interpret non-offending parameters even when the overall solution is improper.

For present purposes, we classified solutions having any parameter estimates with excessively large SEs as improper even when the solution was not otherwise improper (i.e., there were no out-of-range parameter estimates). Whereas this strategy clearly follows from van Driel's (1978) theoretical work, it does not seem to be incorporated into typical practice. However, the results of the present investigation showed that offending (i.e., those with large SEs) and offending-related parameter estimates in this classification were systematically biased. Marsh et al. (1992) reanalyzed previously published data and showed that substantive interpretations based on a "large SE proper" solution (with multiple parameters with large SEs) were unwarranted for those data. More specifically, they demonstrated that even though the results of separate group analyses for two groups appeared to differ substantially, there was good support for the factorial invariance of solutions across the two groups. They argued that even though the separate group solutions appeared to be very different, the large SE solution was so unstable that the conclusion was unwarranted. When invariance constraints were imposed, the common group solution differed substantially from the original solution for the large SE group, but was very similar to the solution from the stable group. Their results suggest that the solution for the large SE group was so unstable that a wide variety of solutions with possibly very different interpretations were not distinguishable in terms of goodness of fit. Consistent with the implications of the present investigation that large-SE-proper solutions should be classified as improper and treated with appropriate caution, the authors recommended that a systematic evaluation of the size of SEs should be incorporated into the evaluation of CFA models.

Goodness of Fit

The original intent of the study was to focus on goodness of fit only as one basis for comparing proper and improper solutions. However, the finding that the apparent fit of true models varied systematically with p/f required further attention. In particular, a superficial -- and erroneous -- interpretation of these findings would argue that researchers should design studies with only two or three indicators per factor -- a suggestion that runs counter to our recommendations that more indicators are better. The fallacy of this interpretation is obvious in

the present investigation because all the approximating models are known to be true. Hence, this apparent decline in fit associated with larger p/f must reflect problems in the standards used to evaluate fit rather than misspecification in the approximating model. In practice, however, researchers would have no way of knowing that the seemingly poorer fit associated with larger p/f actually reflected a bias in the χ^2 statistic. It also follows that this systematic bias in the χ^2 must also affect the behavior on the wide variety of subjective indices of fit that can be expressed as a function of χ^2. Although not presented as part of this study, we found similar effects of p/f, N, and their interaction on the non-normed fit index, the relative noncentrality index, noncentrality, and other subjective indices evaluated in our previous research (e.g., Marsh & Balla, 1994; Marsh et al., 1988; McDonald & Marsh, 1990). Whereas this pattern of results in not of central interest to the present investigation, it apparently has important implications for the typical interpretation of goodness of fit. Because of the importance placed on evaluating the fit of one model against a fixed standard (e.g., the obtained χ^2 being less than its critical value, or incremental fit indices being greater than .90) this bias would naturally extinguish the otherwise desirable strategy of using larger numbers of indicators. This may explain in part why so many published CFA studies are based on p/f = 2 or 3.

Our research is not the only Monte Carlo CFA study in this area to find a systematic relation between the apparent fit of true models and p/f that varies with N (see earlier summaries of Anderson & Gerbing, 1984; Boomsma, 1982; Bearden et al., 1982; Ding, 1993; Ding, et al., in press; Gerbing & Anderson , 1987;. Velicer & Fava, 1987). However, this finding seems not to have been emphasized in previous research and appears not to have influenced typical practice in CFA research. Commenting on this trend in their earlier research that was "not discussed or explained," Gerbing and Anderson (1993, p. 150) noted that the fit of true models declined as the "number of factors in the model, or the number of indicators per factor, increased." They suggested that the explanation may be related to parsimony in that models with fewer indicators have fewer df, "leaving more 'room to maneuver' the parameter estimates so as to minimize the fit function" (p. 50). Alternatively, it may suffice to simply conclude that the speed with which the χ^2 test statistic approaches its asymptotic behavior with increasing N is slower when there are more indicators. Whereas an explanation for this breakdown in the asymptotic behavior the χ^2 test statistic as a function of the number of indicators clearly warrants further research, the implications of this finding are particularly important in the present investigation. Due to this apparent bias in the χ^2 test statistic and goodness of indices that are a function of χ^2, a superficial interpretation of fit statistics would suggest that using fewer indicators is better even though other characteristics considered here (e.g., convergence to proper solutions, interpretability, accuracy and sampling variation in parameter estimates, factor reliability) argue that using more indicators is better.

One anonymous reviewer offered a radical summary of this problem. The reviewer emphasized emphatically that the problems associated with the χ^2 test statistic have long been recognized (e.g., Zwick & Velicer, 1986) but have been largely ignored in structural modeling practice, leading this reviewer to conclude that: "There are two obvious requirements here: (a) The chi-square test statistic should be deleted from all software programs, and (b) Goodness of fit indices that are not a function of the chi-square statistic need to be

developed." Although this concern was not the major focus of the present investigation and the scope of the design is not sufficiently broad to warrant such extreme conclusions, the results reported here are not inconsistent with these recommendations and are consistent with results noted in other research.

This apparent bias in the interpretation of the traditional χ^2 test statistic provides a dilemma for the applied researcher. It may be possible to devise adjusted test statistics or the use of alternative distributions to more appropriately calibrate p-values. Whereas researchers have developed more robust test statistics such as the Satorra-Bentler robust χ^2 test statistic in EQS (Bentler, 1989; Chou & Bentler, 1995), these alternative test statistics have focused on problems associated with non-normal data that may not be very helpful in the present situation (since all data considered here are from normally distributed populations). An alternative approach might be to use bootstrapping techniques where repeated samples with replacement drawn from the original sample are used to estimate an empirical sampling distribution (Bollen & Stine, 1992). Although the bootstrap estimates of the chi-square distribution presented by Bollen and Stine are encouraging, bootstrapping has not been studied sufficiently in CFA research to recommend its routine application. Until more appropriate test statistics are developed, however, a more expedient alternative might be to place more emphasis on comparisons of the relative fit of multiple competing models of the same data -- particularly a priori sets of nested or partially nested models (e.g., models developed to test MTMM data). This emphasis on relative fit substantially reduces the reliance on possibly inappropriate indicators of the absolute fit of any one model. Although the resolution of how to compensate for apparent biases in the traditional χ^2 test statistic is clearly beyond the scope of the present investigation, it is an important area for further research

Parcels .

We also explored the efficacy of using parcels constructed from the 12 items per factor instead of the 12 items as the starting point for CFAs. Solutions based on 2, 3, 4, or 6 parcels performed somewhat better than the corresponding solution based on 2, 3, 4, or 6 items, although trends observed in the item solutions were also evident in the parcel solutions. The more relevant comparison, however, was between solutions based on parcels constructed from the 12 items and the 12-item solutions. The 12-item solutions performed better than the parcel solutions in some cells in that they were more likely to converge to fully proper solutions. However, these differences were only evident when N or the number of parcels was small. The comparison of parameter estimates in the item and parcel solutions was complicated by the change in metric of parcel and item solutions. However, the equivalence of the factor reliability estimates across the parcel and 12-item solutions suggested that there were no particular advantages in using parcels. Consistent with this finding, the size and variability of correlations among the factors did not differ in the parcel and 12-item solutions. Particularly in studies where the substantive emphasis is on the structural aspect of the solution involving relations among the latent constructs, results based on parcels and items seem to be equivalent so long as the number of parcels is reasonable (i.e., at least 3) and the solutions are fully proper.

These results suggest that the number of indicators used to infer a latent construct is more important than whether latent constructs are inferred with items or parcels. In relation to our question of whether or not it

is better to parcel, it seems that it does not make much difference so long as the number of parcels and N is adequate. However, if the N is small or the number of items is not sufficient to construct at least 3 or 4 parcels per factor, then it may be better to conduct analyses at the item level. There are, of course, many other considerations that may impact on the efficacy of parcels (e.g., the use of items with categorical, non-normal response distributions) that are beyond the scope of the present investigation. Thus, for example, scores based on parcels are more likely to be normally distributed than scores based on individual items so that this might be a potential advantage of using parcels instead of items. Also, parcels may be used to "hide" -- wittingly or unwittingly -- unique covariance associated with idiosyncratic characteristics of the items (e.g., the use of positively and negatively worded items), although the desirability of this practice requires further consideration. In the present investigation, however, results based on the parcel solutions provide additional support for our "more is better" conclusion. Having more items to begin with is better whether analyses are done at the item or parcel level, item solutions are somewhat better behaved than parcel solutions, and solutions based on more parcels are somewhat better behaved than solutions based on fewer parcels. As emphasized by an anonymous reviewer, one practical recommendation is that if a researcher obtains an improper solution using parcels, then the use of more parcels per factor or conducting analyses at the item level may resolve the problem.

There may, however, appear to be other advantages in analyzing parcel scores instead of item scores that may be more illusory than real. For purposes of ease of interpretation, researchers often present fully standardized parameter estimates instead of unstandardized estimates. In this case, factor loadings based on parcel scores will be systematically higher than those based on item scores -- substantially so if parcels are based on many items. Larger factor loadings are typically interpreted to mean that the psychometric properties of the underlying measurement instrument are better. However, because both the parcel and item solutions are based on the same set of items, this typical interpretation of the difference in factor loading is inappropriate. The apparent bias in the traditional χ^2 test statistic reported here and elsewhere results in a more subtle illusory advantage of parcel solutions. This is clear in the present investigation because both the parcel and item solutions are based on "true" models. However, even in this ideal situation, the extent of the illusory advantage varies with other characteristics such as N and the juxtaposition between the number of items and parcels. In applied settings with real data, the situation is likely to be even more complicated. Because the illusoriness of these apparent advantages of parcel analyses are not widely recognized and not easily identified in applied research, the preference for the use of parcels is likely to continue. However, because our results suggest that there may be few disadvantages in using parcel scores instead of item scores, the implications of this potential bias in favor of parcels may not be too serious. As emphasized earlier, there is a clear need for further research comparing the advantages and disadvantages under a much wider set of conditions than considered here.

Two-indicator Solutions.

Applied researchers are typically recommended to use $p/f \geq 3$ (e.g., Bollen, 1989), but many applied studies use $p/f = 2$. Whereas our results generally support the typical recommendation, the pattern of results is somewhat more complicated than anticipated. In some respects the disadvantages of using only two indicators

are mitigated by a very large N. However, there were additional problems for solutions with p/f = 2, some of which are not well documented in the literature. It is well known that solutions with p/f=2 are not identified when the latent factors are uncorrelated and are likely to be very unstable when the population correlations are small (Anderson & Rubin, 1956; Bollen, 1989). Because of sampling fluctuations, associated problems with empirical estimation are likely to be exacerbated when N is small. For this reason, two-indicator solutions should be avoided. This well-known feature of two-indicator solutions also points to a potentially important limitation in our study in which factor correlations were fixed at a relatively low value of .3. Many more problems would have been likely if smaller factor correlations had been specified -- particularly if factor correlations approached zero -- whereas even fewer problems would have been likely if larger factor correlations had been specified.

The results of the present investigation also suggest other concerns with two-indicator solutions that are not so well known. Summaries of goodness of fit may be misleading in that χ^2 test and associated p-values are positively biased (i.e., apparent fit is better than it should be; Tables 3 and 6). At least in the present investigation, parameter estimates are also somewhat biased in relation to known population values in that observed factor loadings are too large, uniquenesses are too small, and factor correlations are too large (Table 4). Also, factor reliability estimates tend to be too large. These more subtle problems are not necessarily inconsistent with results of previous research (e.g., Bearden et al., 1982; Gerbing & Anderson, 1987), but they seem not to have been emphasized. These conclusions about p/f =2 clearly warrant further consideration to test the obvious limitations in the generalizability of our study (e.g., all factor correlations were .3) , but the pattern of results supports the recommendation that researchers should not consider p/f = 2.

Limitations and Directions For Further Research

Specific recommendations about the minimum appropriate N and p/f and the use of parcels or items based on the present investigation are necessarily idiosyncratic to this particular study. It is unlikely that these particular values will generalize to other conditions and we have already suggested some variables that may influence these results (e.g., size of factor loadings and factor correlations, degree of misspecification, complexity of factor structures). However, we predict that there will be continued support for our "more is better" conclusion in terms of both N and number of indicators.

Our study considered a much wider variety of p/fs than is typically considered, but it is useful to speculate about potential limitations to our claim that more is better. It might seem logical, for example, that there is some limit on the number of variables that can be reasonably fit when N is small. One possibly critical barrier is where the number of estimated parameters exceeds N (t > N). LISREL (Joreskog & Sorbom, 1993) provides users with an explicit warning that results should be interpreted cautiously when this occurs. In our study, this possibly critical barrier was only crossed for solutions based on p/f=12 (number of estimated parameters = 75) for N=50. However, results summarized here and more detailed inspections of these solutions suggested no dire consequences. These solutions seemed better behaved than p/f = 6 solutions (with 39 estimated parameters) that did not cross this barrier and there were no apparent discontinuities in the trends

observed across all five p/f ratios based on N=50 and N=100. In order to further explore this conclusion, we conducted supplemental simulations for p/f=12 for Ns between 70 and 80 (i.e., Ns slightly above and slightly below the barrier). Again we found no discontinuities in the pattern of results (convergence to proper solutions, fit statistics, and parameter estimates). Although further research is clearly warranted, these supplemental analyses suggest that there may be nothing special about crossing the n=t barrier despite LISREL's dire warning! These supplemental results also support the generality of our conclusions that more is better.

An even more extreme barrier is where N=p. In the present investigation we did not cross this barrier in that the most extreme case we considered was N=50 and p=36. In order to explore this barrier we simulated additional data for p/f=12 (i.e., p = 36) and N=25. For these solutions LISREL appropriately warned us that the input matrix was not positive definite and invoked a default "ridge" option (Joreskog & Sorbom, 1993). In this ridge estimation a constant (.001 by default) multiplied by the diagonal of S is added to S in order to create a positive-definite S as required for maximum likelihood method of estimation. Again we found that the solutions behaved as expected in comparison to solutions with larger Ns. Most (99%) solutions converged to a proper solution and the mean parameter estimates approximated the population values. Although the standard errors of estimates were large (about .2), they were only marginally larger than those based on N=50 and p/f=12. Whereas the p-values for the χ^2 test statistic were noticeably poorer, this apparently represented problems in the asymptotic behavior of the χ^2 discussed earlier. We then simulated data for N=25 for p/f = 2, 3, 4, and 6. Again we found that for a fixed N (N=25), the behavior of the solutions steadily improved with increasing p/f. Our surprising success at N=25 prompted us to push the limits of generalizability even further by examining solutions at N=10 for p/f = 12. Here we did find that the behavior of the solutions deteriorated dramatically in that 20% of the solutions failed to converge at all and only 5% of the solutions were fully proper. However, when we also explored further simulations with N=10 and with p/f = 2, 3, 4, and 6 we found that the performance was even poorer. Hence, not even these analyses based on a completely unrealistic N=10 seemed to be inconsistent with our claim that more indicators is better, and the claim was apparently supported by our supplemental analyses with N=25. It should be emphasized that we do not recommend using such small N and only pursued these analyses to tests the limits of our conclusions. However, if a situation dictates that researchers use unacceptably small Ns -- say 50 or less -- it is better to have a large number of indicators per factor.

Given that we began this study with an apparently counter-intuitive proposal, the results of this investigation and our supplemental analyses provided surprising good support for our conclusion that more is better. Within the obvious limitations of this study, we showed that for a fixed N solutions with more indicators per factor (up to p/f=12) performed better than solutions with smaller p/fs. Furthermore, the advantages of having more indicators were greater when N was small. The solutions were very well behaved if either p/f was large and N was at least moderate or N was large and p/f was at least moderate. In order to avoid any possible misinterpretation, we emphasize again that CFAs should be conducted with large Ns. CFAs based on N<100 or even N<200 should be avoided and our general recommendation is that more is never too much. It would be

inappropriate to cite our study as a justification for using small N. However, if circumstances dictate that N must be small despite our warnings, then it is better to have a large number of indicators per factor. Large p/f can compensate to some extent for small Ns and vice-versa. Whereas the development of a more rigorous mathematical basis for explaining this observation is beyond the scope of the present investigation, our results suggest that this may be a reasonable direction for further research. The results of the present investigation do, however, have a very clear take-home message to the applied researcher that is not evident in typical practice: use more indicators per factor because more is never too much.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In Proceedings of the Third Berkeley Symposium (Vol. 5). Berkeley, CA: University of California Press.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. Journal of Marketing Research, 19, 425-430.

Bentler, P. M. (1989). EQS: Structural equations program manual. Los Angles: BMDP Statistical Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 68, 588-606.

Bentler, P. M., & Jamshidian, M. (1994). Gramian matrices in covariance structure models. Applied Psychological Measurement, 18, 79-94.

Bollen, K. A. (1989). Structural equations with latent variables. NY: John Wiley & Sons.

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 1-9). Newbury Park, CA: Sage.

Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 111-135). Newbury Park, CA: Sage.

Boomsma, A. (1982). Robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskog & H. Wold (Eds.) Systems under indirect observation: Causality, structure, prediction (Part I) (pp. 149-173). Amsterdam: North-Holland.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika, 50, 229-242.

Cattell, R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. New York: Plenum.

Chou, C-P, & Bentler, P. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), Structural equation modeling: Concepts, issues and applications (pp. 111-135). Thousand Oaks, CA: Sage.

Comrey, A. L. (1970). The Comrey Personality Scales. San Diego, CA: Edits Publishers.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. Journal of Consulting and Clinical Psychology, 56,

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. Psychological Bulletin, 105, 317-327.

Ding, L. (1993). The effects of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices. An unpublished thesis, University of Rhode Island.

Ding, L., Velicer, W. F., & Harlow, L. L. (in press). The effects of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices. Structural Equation Modeling, xxx-xxx.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. Psychological Bulletin, 101, 126-135.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecification. Psychometrika, 52, 99-111.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 40-65). Newbury Park, CA: Sage.

BEST COPY AVAILABLE

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275.

Joreskog, K. G., & Sorbom, D. (1988). LISREL 7 - A guide to the program and applications (2nd ed.). Chicago, Illinois: SPSS.

Joreskog, K. G., & Sorbom, D. (1993). LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International.

Marsh, H. W (1988). Self Description Questionnaire I: A manual and a research monograph. San Antonio, TX: The Psychological Corporation.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait- multimethod data: Many problems and a few solutions. Applied Psychological Measurement, 13, 335-361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. Applied Psychological Measurement, 15, 47-70.

Marsh, H. W., & Balla, J. R. (1994). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size and model complexity. Quality & Quantity, 28, 185-217.

Marsh, H. W., Balla, J. R., McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Marsh, H. W., Byrne, B. M., & Craven, R. (1992). Overcoming problems in confirmatory factor analysis of MTMM data: The correlated uniqueness model and factorial invariance. Multivariate Behavioral Research, 27, 489-507.

Marsh, H. W., & O'Niell, R. (1984). Self Description Questionnaire III (SDQIII): The construct validity of multidimensional self-concept ratings by late-adolescents. Journal of Educational Measurement, 21, 153-174.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. Psychological Bulletin, 107(2), 247-255.

Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

SPSS (1993). SPSS Reference Guide. Chicago: SPSS

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243.

Velicer, W. F., & Fava, L. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. Multivariate Behavioral Research, 22, 193-209.

Velicer, W. F., & Fava, L. L. (1994). The effect of variable and subject sampling on component, image, and factor analysis. (in review).

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. Multivariate Behavioral Research, 25, 1-28.

Wothke, W. (1993). Nonpositive definite matrices in structural equation modeling. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 256-293). Newbury Park, CA: Sage.

Zwick, W. R., & Velicer, W. F. (1986). A comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

Figure Captions

Figure 1. An example of one population target model (p/f = 3) used to generate the simulated data.

Figure 2. Box-plots for Goodness of Fit as a function of the number of indicators per factor and the convergence behavior: χ^2/df ratio and p-value associated with the χ^2 test of statistical significance.

Figure 3. Box-plots for parameter estimates for fully proper solutions as a function of the number of indicators per factor and sample size: factor loadings and factor correlations.

Figure 4. Box-plots for Goodness of Fit as a function of the number of indicators per factor and sample size: χ^2/df ratio and p-value associated with the χ^2 test of statistical significance.

Table 1

Percentages of Different Convergence Behaviour by Sample Size and Number of Indicators/Parcels per Factor

Number of Indicators	Solution Type	Item Solutions: Samples Size 50	100	200	400	1000	Parcel Solutions: Sample Size 50	100	200	400	1000
2	Proper	13.6	32.8	55.6	82.4	93.0	15.6	35.3	69.0	92.0	99.0
	Boundary	2.8	2.2	1.0			30.7	34.0	19.6	5.2	
	Non-Boundary		.1				2.7	1.1			
	SE large proper	3.3	6.0	8.6	8.8	7.0	.4	1.4	1.6	1.6	1.0
	SE large improper	23.8	25.8	22.2	6.4		37.7	26.0	9.6	1.2	
	Nonconvergent	56.6	33.1	12.6	2.4		13.0	2.2	.2		
3	Proper	54.8	85.4	97.8	100.0	99.0	97.9	100.0	100.0	100.0	100.0
	Boundary	2.0	.2				1.7				
	Non-Boundary	.1					.1				
	SE large proper	12.0	10.7	2.2		1.0	.2				
	SE large improper	20.4	2.8				.2				
	Nonconvergent	10.8	.9								
4	Proper	86.5	99.1	99.6	100.0	100.0	99.8	100.0	100.0	100.0	100.0
	Boundary	1.2	.1				.2				
	SE large proper	5.5	.7	.4			.1				
	SE large improper	4.7	.1								
	Nonconvergent	2.2									
6	Proper	99.6	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0
	SE large proper	.2					.1				
	SE large improper	.2									
	Nonconvergent	.1									
12	Proper	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note. Percentages are based on 2500 replicates of N=50, 1000 replicates of N=100 cases, 500 replicates of N=200, 250 replicates of N=400, and 100 replicates of N=1000. Each solution was classified as fully proper or as falling into one the categories of improper solution. For purposes of Study 3 only, the 12 items/factor in the 12 item solution were used to construct parcels (e.g., the 12 items were divided into three parcels of four items each and solutions based on these three indicators were evaluated).

Table 2
Parameter Estimates by Convergence Behaviour and Number of Indicators

Solution Behavior	Factor Loading: Proper Mean SD	Factor Loading: Non-Offend Mean SD	Factor Loading: Offend-Related Mean SD	Factor Loading: Offend Mean SD	Uniqueness: Proper Mean SD	Uniqueness: Non-Offend Mean SD	Uniqueness: Offend Related Mean SD	Uniqueness: Offend Mean SD	Factor Correlation: Proper Mean SD	Factor Correlation: Non-Offend Mean SD	Factor Correlation: Offend Related Mean SD	Factor Correlation: Offend Mean SD
Two Indicators per Factor												
Proper	.61 .23 (2023)				.59 .24 (2034)				.41 .29 (1017)			
Boundary		.60 .18 (290)	.46 .09 (62)	1.07 .10 (62)		.61 .23 (290)	.82 .19 (62)	-.14 .11 (62)		.47 .26 (74)	.37 .20 (121)	1.11 .11 (12)
SE Large (Proper)		.55 .27 (342)	.29 .23 (77)	.74 .52 (79)		.61 .24 (342)	.87 .20 (77)	.25 .23 (79)		.49 .30 (88)	.16 .25 (141)	.32 .65 (20)
SE Large (Improper)		.55 .30 (1846)	.26 .18 (853)	1.25 1.03 (865)		.60 .24 (1846)	.91 .18 (853)	-1.58 2.78 (865)		.42 .32 (327)	.15 .19 (1334)	.47 .68 (121)
Nonconvergent		.49 .38 (2508)	.10 .18 (2112)	1.85 5.53 (3870)		.61 .25 (2508)	.95 .20 (2112)	-32.98 50.50 (3870)		.37 .35 (329)	.05 .16 (1396)	-.16 38.57 (2520)
Three Indicators per Factor												
Proper	.61 .20 (8214)				.61 .19 (7440)				.34 .23 (4107)			
Boundary		.61 .20 (206)	.77 .14 (70)	.99 .10 (30)		.61 .20 (206)	.77 .14 (70)	-.06 .07 (30)		.37 .21 (52)	.31 .24 (98)	1.05 .01 (3)
SE Large (Proper)		.59 .19 (1244)	.36 .12 (378)	.94 .11 (172)		.61 .20 (1244)	.80 .16 (378)	.23 .15 (172)		.30 .24 (342)	.20 .20 (532)	.38 .56 (23)
SE Large (Improper)		.59 .20 (1782)	.29 .14 (832)	1.34 .72 (440)		.60 .21 (1782)	.84 .18 (832)	-1.28 3.26 (440)		.33 .26 (381)	.14 .21 (1065)	.30 .69 (81)
Nonconvergent		.60 .18 (704)	.11 .16 (429)	2.50 5.21 (493)		.62 .22 (704)	.87 .17 (429)	-32.42 48.54 (493)		.31 .24 (134)	.08 .17 (149)	2.21 25.27 (530)
Four Indicators per Factor												
Proper	.60 .17 (12978)				.62 .18 (12978)				.30 .21 (6489)			
Boundary		.58 .17 (120)	.44 .13 (38)	.95 .12 (16)		.61 .19 (120)	.73 .17 (38)	-.09 .07 (16)		.27 .19 (31)	.26 .21 (54)	1.11 .01 (2)
SE Large (Proper)		.59 .17 (568)	.36 .13 (185)	.89 .21 (69)		.61 .17 (568)	.76 .16 (185)	.29 .20 (69)		.25 .21 (163)	.16 .20 (247)	.45 (1)
SE Large (Improper)		.59 .16 (452)	.27 .14 (185)	1.24 .35 (65)		.62 .18 (452)	.80 .17 (185)	-.63 1.15 (65)		.29 .22 (108)	.13 .19 (231)	.43 1.09 (12)
Nonconvergent		.60 .17 (174)	.07 .13 (71)	3.05 5.13 (79)		.62 .19 (174)	.91 .17 (71)	-34.35 53.72 (79)		.31 .19 (41)	.07 .13 (21)	2.11 22.12 (100)

Note. Six sample indicators (two from each factor) and three correlations were used in the calculation of average loadings, uniqueness, and factor correlations. However, the entire set of parameter estimates was used to classify each solution as proper or one of the classifications of improper solution. For fully proper solutions, all parameter estimates are proper, but for improper solutions each parameter estimate was classified as offending, offending-related, or non-offending (not associated with a factor having an offending parameter estimate). The number of parameter estimates used in the calculation of each means and SD were shown in brackets. Population values from the generating model (see Figure 2) are .60 for factor loadings, .30 for factor correlations and .64 for uniquenesses.

Table 3

Goodness of Fit For Proper and Improper Solutions: Tests of statistical significance Conducted Separately For Each Solution Type.

No. of Indicators / Solution Type	N	Chi/df Mean	Chi/df SD	p-value Mean	p-value SD
2 Indicators/Factor					
Proper	339	.788[a]	.421	.605[a]	.258
Boundary	69	.801[a]	.457	.601[a]	.273
SE large proper	83	.724[a]	.393	.644[a]	.255
SE large improper	594	.754[a]	.443	.630[a]	.266
Nonconvergent	1415	1.075[b]	.903	.486[b]	.282
F(4,2495) =		27.31***		38.25***	
3 Indicators/Factor					
Proper	1369	1.038[a]	.300	.466[a]	.291
Boundary	51	1.072[ab]	.315	.438[ab]	.293
SE large proper	299	1.006[a]	.306	.497[a]	.297
SE large improper	509	1.092[ab]	.301	.411[ab]	.285
Nonconvergent	271	1.165[b]	.327	.351[b]	.273
F(4,2495) =		13.68***		12.18***	
4 Indicators/Factor					
Proper	2163	1.109[a]	.224	.360[a]	.281
Boundary	29	1.179[a]	.233	.279[a]	.268
SE large proper	137	1.129[a]	.239	.344[a]	.289
SE large improper	117	1.146[a]	.220	.307[a]	.263
Nonconvergent	54	1.202[a]	.208	.237[a]	.223
F(4,2495) =		3.68**		4.04**	

Note. There was only one improper solution falling into the non-boundary classification, and so this classification is not presented. A oneway ANOVA was conducted separately for each index, followed by a post-hoc test of pairwise comparisons. Means not statistically different from one another ($p < .05$) are indicated by the same superscript.

*p <.05, **p < .01, ***p < .001.

Table 4
Parmeters and Factor Reliability Estimates by Sample Size and Number of Indicators

	Item Solutions												Parcel Solutions											
	Fac Load		Unique		Fac Corr		Reliability		Chi/df		p-value		Fac Load		Unique		Fac Corr		Reliability		Chi/df		p-value	
Number of Indicators	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Sample Size=50																								
2	.611	.235	.594	.236	.411	.288	.564	.127	.788	.421	.605	.258	.606	.083	.104	.055	.382	.142	.874	.035	.874	.497	.561	.272
3	.615	.171	.609	.197	.336	.228	.649	.085	1.038	.300	.466	.291	.600	.090	.156	.052	.297	.158	.871	.033	1.099	.330	.414	.296
4	.603	.166	.615	.179	.305	.211	.694	.071	1.109	.224	.360	.281	.599	.096	.209	.059	.296	.157	.870	.032	1.133	.227	.328	.278
6	.599	.151	.625	.160	.296	.184	.769	.054	1.194	.146	.154	.196	.600	.109	.311	.078	.296	.157	.870	.029	1.196	.150	.154	.200
12	.599	.146	.624	.143	.291	.160	.870	.029	1.476	.091	.000	.000	.599	.146	.624	.143	.291	.160	.870	.029	1.476	.091	.000	.000
Sample Size=100																								
2	.612	.190	.597	.206	.365	.189	.569	.093	.823	.465	.587	.279	.604	.067	.104	.052	.332	.100	.874	.027	.933	.528	.532	.284
3	.606	.129	.622	.146	.305	.161	.635	.062	1.032	.296	.470	.293	.601	.063	.158	.036	.298	.109	.871	.023	1.052	.301	.451	.288
4	.599	.114	.631	.124	.299	.142	.694	.050	1.058	.212	.424	.290	.601	.067	.211	.042	.298	.108	.871	.021	1.067	.202	.403	.281
6	.602	.105	.635	.112	.298	.124	.771	.037	1.092	.127	.302	.253	.601	.075	.316	.054	.298	.108	.871	.020	1.092	.137	.309	.268
12	.601	.096	.631	.098	.298	.108	.871	.019	1.171	.069	.026	.064	.601	.096	.631	.098	.298	.108	.871	.019	1.171	.069	.026	.064
Sample Size=200																								
2	.611	.151	.609	.172	.328	.130	.555	.070	.909	.490	.538	.278	.602	.053	.104	.046	.308	.072	.874	.019	.982	.541	.503	.292
3	.602	.091	.632	.102	.295	.110	.628	.046	1.021	.286	.479	.280	.602	.045	.159	.025	.299	.075	.872	.016	1.019	.280	.477	.284
4	.599	.081	.636	.087	.300	.097	.694	.034	1.036	.207	.451	.292	.602	.048	.213	.029	.299	.075	.871	.015	1.034	.201	.452	.290
6	.603	.074	.640	.078	.298	.084	.772	.026	1.051	.127	.385	.274	.602	.053	.318	.038	.299	.074	.871	.014	1.045	.119	.395	.271
12	.602	.068	.635	.068	.299	.074	.872	.014	1.077	.060	.180	.203	.602	.068	.635	.068	.299	.074	.872	.014	1.077	.060	.180	.203
Sample Size=400																								
2	.612	.113	.615	.139	.302	.091	.551	.053	.952	.560	.526	.292	.603	.041	.105	.037	.300	.054	.874	.014	.935	.572	.539	.292
3	.603	.062	.637	.071	.296	.076	.627	.033	1.000	.276	.498	.278	.602	.031	.160	.018	.299	.054	.872	.011	.996	.289	.502	.297
4	.599	.055	.639	.061	.300	.071	.694	.024	1.018	.205	.476	.293	.602	.033	.214	.021	.299	.054	.872	.011	1.030	.197	.455	.281
6	.603	.052	.642	.054	.299	.060	.772	.019	1.029	.131	.435	.295	.603	.038	.319	.027	.299	.054	.872	.010	1.026	.118	.434	.281
12	.602	.047	.637	.048	.299	.054	.872	.010	1.035	.058	.339	.266	.602	.047	.637	.048	.299	.054	.872	.010	1.035	.058	.339	.266
Sample Size=1000																								
2	.606	.067	.630	.080	.299	.057	.539	.029	.969	.581	.517	.321	.603	.026	.106	.024	.299	.033	.872	.009	.942	.528	.522	.292
3	.603	.040	.639	.046	.296	.046	.627	.020	1.025	.325	.488	.301	.602	.020	.160	.011	.299	.033	.872	.007	.974	.289	.521	.294
4	.599	.035	.641	.039	.300	.046	.694	.015	1.037	.217	.457	.299	.602	.021	.214	.014	.299	.033	.872	.007	1.026	.203	.467	.276
6	.603	.033	.643	.034	.299	.035	.772	.012	1.028	.115	.423	.273	.603	.024	.319	.017	.299	.033	.872	.007	1.010	.116	.470	.291
12	.603	.029	.638	.031	.299	.033	.872	.006	1.015	.054	.422	.276	.603	.029	.638	.031	.299	.033	.872	.006	1.015	.054	.422	.276

Note. Values are based only on fully proper solutions (see Table 1). Results for the parcel solutions are considered in Study 3. Population values from the generating model (see Figure 2) are .60 for factor loadings, .30 for factor correlations and .64 for uniquenesses.

BEST COPY AVAILABLE

Table 5

Effects of Sample Size and Number of Indicators/Parcels on Parameter Estimates, Factor Reliability and Goodness of Fit Indexes

	Item Solutions						Parcel Solutions					
	Number of Indicator		Sample Size		N Indicator X Size		Number of Indicator		Sample Size		N Indicator X Size	
	E	r	E	r	E	r	E	r	E	r	E	r
Parameter Estimates												
Factor Loading	.02	-.01	.01	-.01	.02	.01	.01	-.00	.01	.01	.01	.00
Uniqueness	.04	.03	.05	.05	.02	-.01	.66	.64	.01	.01	.01	.01
Factor Correlation	.06	-.04	.06	-.06	.08	.06	.05	-.03	.04	-.03	.07	.04
Parameter Variability												
Factor Loading	.20	-.17	.29	-.29	.07	.02	.11	.08	.30	-.29	.11	-.10
Uniqueness	.23	-.19	.30	-.30	.06	.03	.18	.11	.28	-.28	.15	-.13
Factor Correlation	.12	-.11	.35	-.34	.08	.07	.00	.00	.32	-.32	.01	-.00
Factor Reliability	.64	.64	.02	-.02	.04	.03	.03	.00	.00	.00	.00	.00
Goodness of Fit Indexes												
Chi/df	.20	.19	.16	-.13	.25	-.22	.16	.15	.18	-.17	.21	-.17
p value	.27	-.27	.14	.14	.18	.16	.25	-.25	.17	.17	.16	.13

Note. For each estimate a 5 (levels of N) X 5 (levels of p/FN) was conducted. Effect size for each main and interaction effect is summarized by eta [E; $(SS_{effect}/SS_{total})^{1/2}$] and the linear effect (r) of log N, log p/N, and their interaction. Values are based only on fully proper solutions (see Table 1). Results for the parcel solutions are considered in Study 3.